DamID Profiling of CTCF and TEAD3 Binding within Chromosome 9p21

Research Thesis

Presented in Partial Fulfillment of the Requirements for Graduation “with Research Distinction in Molecular Genetics” in the Undergraduate Colleges of The Ohio State University

By
Adam Sychla

The Ohio State University
May 2018

Project Advisor: Dr. Christin Burd
Department of Molecular Genetics, Department of Cancer Biology and Genetics

Thesis Committee: Dr. Sharon Amacher, Dr. Christin Burd, Dr. Paul Herman, Dr. F. Kay Huebner

Abstract

The progression of benign tumor growth is often restricted by oncogene-induced senescence (OIS), a state of irreversible cell cycle arrest that frequently involves upregulation of the $p16^{INK4a}$ tumor suppressor. Senescence is ultimately controlled by the accumulation of $p16^{INK4a}$, but the intermediary steps which cause $p16^{INK4a}$ transcriptional upregulation are unknown. Here, I employ the DNA adenine methyltransferase identification (DamID) assay to investigate protein binding to DNA elements within the *INK/ARF* locus, which encodes $p16^{INK4a}$. In DamID, an *E. coli* Dam methylase is fused to DNA-binding factors of interest. Interaction of the Dam-fused DNA binding factor with DNA causes local adenosine methylation. Patterns of DNA binding can then be investigated using a methylation-sensitive restriction enzyme digest followed by qPCR for the region of interest. I selected putative *INK/ARF* binding proteins and made three fusion constructs: Dam-CTCF, Dam-TEAD3, and Dam-SMAD3. I mapped the steady-state binding patterns for each of these fusion proteins across the 90kb *INK/ARF* regulatory region. In this way, I quantified the usage of several known binding sites and identified novel interactions. With this knowledge, we can now investigate how oncogene activation remodels DNA-binding factor interactions to engage, and eventually subvert, the tumor suppressive functions of the *INK/ARF* locus.

Introduction

$p15^{INK4b}$, $p16^{INK4a}$ and $p14^{ARF}$ are encoded by the *INK/ARF* locus located on chromosome 9p21.3. These proteins play an integral role in tumorsuppression and are functionally inactivated in most cancers [1]. The locus is organized into two genes, *CDKN2A* and *CDKN2B*. *CDKN2A* consists of four exons and generates two distinct mRNA products from separate promoters, one encoding $p16^{INK4a}$ and one encoding $p14^{ARF}$ each with a unique first exon (Figure 1).While the last two exons of *CDKN2A*are shared by both mRNAs, they are translated in different open reading frames (ORFs). In this way,*CDKN2A* encodes two functionally different tumor suppressors without protein homology. Upstream of the $p14^{ARF}$ promoter, lies *CDKN2B* which contains two exons that are spliced together to producethe $p15^{INK4b}$ mRNA (Figure 1).

Protein products of the *INK/ARF* locus control the cell cycle through two primary pathways. $p14^{ARF}$ stabilizes the activity of p53, a tumor suppressor that controls cell cycle arrest, initiates apoptosis, and engages mechanisms involved in DNA damage repair [1, 2]. $p14^{ARF}$ achieves this through inhibition of the E3 ubiquitin-protein ligase, MDM2,and this action leads to p53 stabilization (Figure 2A)[3]. $p15^{INK4b}$ and $p16^{INK4a}$ are involved in the Retinoblastoma (Rb) tumor suppressor pathway [4]. Active E2F transcription factors upregulate a collection of genes involved in the G_1- to S-phase cell cycle transition. Rb binding inhibits E2F activity, but this interaction is lost when Rb is phosphorylated by activated Cyclin Dependent Kinase 4 (CDK4) and/or Cyclin Dependent Kinase 6 (CDK6)(Figure 2B)[6]. CDK4/6 are activated by binding with Cyclin D [4]. $p15^{INK4b}$ and $p16^{INK4a}$ disrupt the interactions between CDK4/6 and Cyclin D thereby preventing RB phosphorylation and subsequent cell cycle progression (Figure 2C)[3, 4]. In response to stressors, such as telomere shortening, DNA damage, or oncogene activation, $p16^{INK4a}$ is upregulatedand cell cyclearrest ensues [1, 5, 6]. This state is normally reversible, but long-term upregulation of these proteins results in cellular senescence, a state of permanent cell cycle arrest. This mechanism is initiated in response to oncogene activation and thus acts as a check against uninhibited growth, termed oncogene-induced senescence (OIS)[1, 5]. Functional inactivation of $p16^{INK4a}$ allows cells to bypass OIS and progress to form tumors [1, 5].

INK/ARF transcription is controlled by a large regulatory region spanning over 90kb on chromosome 9p21. Multiple genome wide association studies (GWAS) have linked single nucleotide polymorphisms (SNPs) in this region to increased risk of cancer and cardiovascular disease (CAD)[7-9]. Some of the CAD-associated SNPs occur in a putative enhancer region that we dubbed the polymorphic regulatory region (PRR). The PRR is located 70 kb to 100kb upstream from the *INK/ARF* locus and was previously shown to enhance luciferase expression *in vitro* in a genotype-dependent manner [10]. In addition to the PRR, a separate study determined that the *INK/ARF* regulatory region contains nine putative enhancer sites dubbed ECAD1-9. Notably, several CAD-associated SNPs are distributed within these enhancer sites [7]. These SNPs could potentially inhibit or enhance binding by unknown *INK/ARF* regulatory factors. The 9p21 interval that spans the *INK/ARF* locus and PRR also contains seven binding sites for CCCTC-Binding Factor (CTCF), a chromatin organizer with insulator and chromatin looping activities[11-14].

To measure the binding of several proteins throughout the *INK/ARF* locus, I made use of the DNA adenine methyltransferase identification (DamID) system. The Dam enzyme, isolated from *E. Coli*, methylates adenine in the DNA sequence, GATC [15-17]. DNA binding proteins fused to Dam lead to preferential methylation near the protein binding location (Figure 3A). Since mammalian cells do not perform adenine methylation, we can quantitatively measure methylation added by the Dam enzyme to identify DNA binding patterns of the protein of interest. To do this, DNA is extracted from cells

expressing the Dam fusion protein and digested with *DpnII*, a methylation-sensitive restriction enzyme that will cut only unmethylated GATC sequences. However, methylated GATC sequences will remain intact (Figure 3B)[15, 16]. Primers are selected to produce an amplicon containing at least one GATC sequence. qPCR is then used to quantify the amount of uncut, methylated DNA which correlates with binding of the fusion protein (Figure 3C). The strong propensity of Dam to methylate GATC sequences leads to background methylation, which must be accounted for through the simultaneous assessment of negative binding control regions [15, 16]. The DamID system has several advantages over the more traditional chromatin immunoprecipitation (ChIP) assays [15]. DamID does not require protein-specific antibodies and can detect transient interactions that might be lost during immunoprecipitation [15]. Further, because Dam acts to methylate any DNA in its proximity, dynamic chromatin loops can also be investigated. Loops will allow for methylation to occur even when the transcription factor binds several kilobases away (Figure 4)[18, 19]. In this way, the DamID assay investigates the relative three-dimensional positioning of regions of DNA and binding proteins.

My investigation focused on using the DamID assay to determine how several putative *INK/ARF* binding proteins interacted with the locus and the 90kb regulatory region. Future investigations will look at changes in 9p21 protein occupancy that may occur during OIS or as a result of polymorphisms in the regulatory region.

Methods

Dam Fusion Constructs

The TEAD3 cDNA was ordered from Transomic while the SMAD3 and CTCF cDNAs were generated from mRNA purified from HEK 293T cells. The cDNA was then amplified by PCR using primers containing homology arms to the lentiviral pLgw-EcoDam plasmid, kindly provided by the van Steensel lab. Using a Gibson cloning reaction (New England Biolabs), each DNA binding factor was inserted in frame with Dam on the N-terminus.

Culture, Transduction, and Collection

Each pLgw-EcoDam plasmid was transfected into HEK 293T cells along with the VSVG envelope protein (pCMV-VSVG) and the packaging vector, pSPAX2 (Figure 5A). Replication-deficient lentiviruses were collected via standard procedure [20].A human immortalized primary melanocyte cell line (hMELT) was a kind gift of the Sellers Lab [21] and was grown in Ham's F12 supplemented with 7.5% FBS, 1% penicillin–streptomycin, 2 mmol/L L-glutamine, 50 ng/ml TPA, 100 µM IBMX, 1 µM Na_3VO_4, and 50 µM dbcAMP. The cells were transduced to create stable cell lines expressing the fusion Dam-proteins (Figure 5B). 500 µg/mL of hygromycin was used to select for transduced cells over the course of 3 days. The lines were maintained with media containing 125 µg/mL of hygromycin. PCR amplification of genomic DNA was used to confirm integration of the lentiviruses.

Cells were plated at two million cells per 10 cm plate and allowed to grow for approximately 24 hours before collection for DamID analysis.DNA was extracted from the hMELT cells that were lysed in lysis buffer (50mM Tris(pH 8.0), 0.1M EDTA (pH 8.0), 0.1M NaCl, and 1% SDS) and purified following a phenol:chloroform clean-up procedure [22]. The DNA was re-suspended in TE Buffer and the concentration analyzed via Qubit assay (Invitrogen). For each sample, 300 ng of DNA was incubated with or without *DpnII*. The digest was incubated at 37 °C for 16 hours. The 300ng of incubated DNA was then diluted to 5ng/µL for qPCR analysis (Figure 5C).

DamID Assay and Analysis

5ng of DNA, 3 µL of 10 ng/µL primer, and 4 µL of SYBR mix (Bio-Rad) were loaded onto a 384 well plate in triplicate. Quantitative polymerase chain reaction (qPCR) was run on the on the genomic DNA of the fusion-containing cell lines with and without *DpnII* digestion (95^{o}C, 10 min; 40x [95^{o}C, 5 sec; 65^{o}C, 30 sec]; melt curve 65^{o}C-95^{o}C, 0.5^{o}C/cycle, 5 sec/cycle). The presence of multiple peaks in the melt curve indicated non-specific binding and was used to remove primer sets. Negative control primer sets with higher C_t values in *DpnII* digested genomic DNA were excluded as well. C_t values that were determined to be outliers by the Dixon Q outlier test at 95% confidence within a triplicate were removed [23, 24]. An average C_t value was then determined for each triplicate. For *DpnII* treated samples, a ΔC_t value was calculated by ΔC_t = C_t *DpnII* treated – average C_t *DpnII* untreated. ΔC_t values for all of the negative control regions were averaged. A $\Delta\Delta C_t$ value was calculated for each sample by the difference of a ΔC_t value and the negative control average ΔC_t value. All data is reported as the triplicate average of $2^{-\Delta\Delta Ct}$. This returns the fold increase in methylation of a triplicate in comparison to the average among negative controls.

Plots were generated to display relative confidence in Dam proximity at a given location. A normal distribution was generated for each primer set, n, with the function: $f_n(x) = C(n) * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $(Equation\ 1)$ where x is the chromosomal base pair value. Centering the GATC sequence sets (µ) to the base pair coordinate of the sequence. σ was chosen to be equal to 1250/3. This value was selected because it would result in 99% of the distribution falling within 1250bp of the Dam enzyme binding site. C(n) was a constant calculated by $C_w(n) = \frac{(\%\ Positive\ )(Average\ \ 2^{-\Delta\Delta Ct}\ \text{values}\ )(N)}{StandardDeviationof\ \ \ 2^{-\Delta\Delta Ct}\ \text{values}}$ $(Equation\ 2)$, $\mathrm{C(n)} = \frac{\mathrm{C_w(n)}}{\mathrm{Max\,(C_w(n))}} * 100$ $(Equation\ 3)$. In this way, C(n) was higher for strong and consistent Dam proximity results, reflecting a higher confidence. A value of f(x) equal to or above fifty was generally considered as evidence for an interaction site. The sum of the normal distributions for a given binding protein was used to visualize Dam methylation patterns.

Primer Selection

For the DamID assay, primer sets had to be designed for each region of interest (Table 1). Similar sets were designed as positive and negative controls for each of the DNA binding factors. Primer sets were developed to amplify regions of about 125bp containing at least one GATC sequence. With GATC sequences occurring on average every 200-300bp in the genome, any region of interest had several to choose from [15]. BLAST was used to remove primer sets that bind to secondary locations in the genome with high affinity [25].

Results

Three DNA-binding factors were selected for investigation: CTCF, SMAD3, and TEAD3. A previous study used ChIP and chromatin conformation capture to identify three CTCF binding sites in the *INK/ARF* locus [11]. These binding sites are located downstream of *p16^{INK4a}* and in the promoter regions of *p15^{INK4b}* and *p14$^{AR\ F}$*(Figure 6)[11]. Further, it was shown that CTCF mediates differential *INK/ARF* chromatin looping structures in somatic, iPS, and senescent cells [11]. A known chromatin organizer involved in insulator activity and chromatin looping, CTCF binding to the *INK/ARF* locus and PRR could potentially reorganize the *INK/ARF* locus to alter gene accessibility [13, 14].

The PRR was shown by luciferase assay to act as a regulatory element and had increased activity with when containing the CAD-associated risk allele [10]. Bioinformatic analysis also indicated that there is a putative SMAD3 binding motif in the PRR [10]. The risk allele disrupts this site suggesting the CAD SNP could alter SMAD3 binding and lead to *INK/ARF* deregulation. Further, SMAD3 is required for the anti-proliferative effects of TGF-β mediated by $p16^{INK4a}$ [10, 26].

It was previously shown that TEAD3 is a positive regulator of the *INK/ARF* locus. In particular, over-expression of the transcription factor led to increased expression of $p16^{INK4a}$ [26]. However, homozygous cell lines for either of two CAD-associated SNPs did not respond to over expression of TEAD3 [26]. TEAD3 was shown by ChIP to bind these SNPs but the risk allele disrupted the binding [26]. Given that TEAD3 has been shown to be an important regulator of the *INK/ARF* locus and that TEAD3 fails to induce $p16^{INK4a}$ with the risk allele, it was chosen to be investigated [26]. Together, these data suggest that CTCF, SMAD3, and TEAD3 are candidates for mediating *INK/ARF* regulation.

SNPs in the PRR are associated with the downregulation of transcripts emanating from the *INK/ARF* locus and other phenotypes including CAD and melanoma [7-9]. In order to investigate binding of CTCF, SMAD3, and TEAD3 to the *INK/ARF* locus, the DamID assay was utilized. The ECAD4 region had been described as having enhancer marks such as H3 monomethylation at lysine 4 (H3K4me1)[7]. A TEAD3 binding site that is lost with the risk allele of a CAD-associated SNP also lies just downstream of this region [27]. For these reasons TEAD3 was chosen to be investigated with primer set ECAD4-1 (Figure 6). A GATC was found within the aforementioned putative SMAD3 binding site. It shared proximity to a putative CTCF binding site in the PRR allowing investigation of several potential modes of regulation through one amplicon. Three primer sets dubbed PRR1, PRR2, and PRR3, were designed to examine this region (Figure 6).

Positive control primer sets were designed for each DNA binding protein. Positive controls primer sets for Smad3, SMAD-PC-1 and SMAD-PC-2, lie within the *Transgelin* promoter where SMAD3 was previously shown to bind [28]. TEAD3 positive control primer sets, TEAD3-PC-1, TEAD3-PC-2, and TEAD3-PC-3, were selected within *Connective Tissue Growth Factor* (*CTGF*), a gene known to be regulated by TEAD3 [29]. I used the UCSC Genome Browser to visualize an ENCODE ChIP-Seq dataset for TEAD3 genomic binding and select specific regions to use as positive control amplicons [27, 30]. Of note, ENCODE data indicated that CTCF had a binding site near the TEAD3 positive controls [27]. In addition to this site, three known CTCF binding sites were investigated within the *INK/ARF* locus. It was previously shown that CTCF binds *CDKN2B* and the $p14^{ARF}$ transcriptional start site and that CTCF is brought into proximity of the $p16^{INK4a}$ transcriptional start site by chromatin looping [11, 12]. Therefore, four CTCF positive controls primer sets were made. The CTCF-PC-1 and CTCF-PC-2 primer sets fall within the *CDKN2B*. CTCF-PC-3 falls in the first exon of $p14^{ARF}$and finally, CTCF-PC-4 lies in the $p16^{INK4a}$ promoter region (Figure 6). My selected TEAD3 and CTCF positive control amplicons indicated binding in my DamID assays (Figures 6, 8A, 9A). However, the SMAD3 positive control primer sets failed to amplify after multiple trials, suggesting that SMAD3 was not binding to these sites in hMELT cells.

Negative controls for DamID were designed using an ENCODE ChIP-Seq dataset for each protein and visualized in the UCSC Genome Browser [27, 30]. Specifically, I searched for regions of the human genome containing no known binding sites for the proteins of interest. Ultimately, five negative control primer sets were selected to be used in my experiments: Chr04-NC-1, Chr06-NC-3, Chr06-

NC-5, Chr06-NC-6, and Chr08-NC-1. These functioned well with no significant evidence of Dam interactions in any of the Dam fusion cell lines and a maximum C(n) value of 12.5 (Figure 8B).

Once working controls were established, DamID analysis was done on each of the three cell lines. The Dam-SMAD3 cell line DNA was not strongly methylated at any of the tested regions within the *INK/ARF* locus or PRR (Figure 9). Interaction peaks that were seen did not exceed the confidence seen in negative controls (Figure 9). The exception was at the *CTGF* promoter region used as the TEAD3 positive control (Figure 8A). While this amplicon peak did rise above negative controls, it was still much smaller than the peaks of the other DNA binding factors in this region. As stated before, I was unable to detect SMAD3 binding at positive control regions (data not shown).

Dam-TEAD3 showed strong evidence of DNA methylation at the positive control region in *CTGF* (Figure 8). Within the *INK/ARF* locus, there was no evidence of interaction with the $p16^{INK4a}$ transcriptional start site, while the $p14^{ARF}$ transcriptional start site showed a potential interaction (Figure 9A). However this signal was not significantly higher than background. There was more evidence for the interaction of TEAD3 with *CDKN2B.* Binding at ECAD4 was also confirmed. A weak TEAD3 binding peak within the PRR is close to that of the negative controls and likely represents background (Figure 9).

The CTCF positive controls in the *INK/ARF* locus showed strong binding peaks indicating that the locus could be used effectively as a control (Figures 9A). Dam-CTCF showed low binding to ECAD4 enhancer region. While there is a CTCF binding site in this region, it also lies about 2kb from the selected primer set and thus does not fall in the usual range of the Dam fusion [14, 15, 12]. In contrast, the PRR and *CTGF* primer sets all showed interactions with high confidence (Figures 8A, 9B). *CTGF* has a known CTCF binding site and thus my results were expected. However, the site at the PRR was not previously shown and therefore, this data confirms a novel binding site for CTCF. Taken together, these data establish DamID as a useful tool for probing the occupancy of putative DNA binding sites within chromosome 9p21.

Discussion

DamID is an alternative to chromatin immunoprecipitation (ChIP) that can be used to investigate protein:DNA interactions [15, 16]. DamID provides several key advantages over ChIP. First, ChIP is reliant on having protein specific antibodies which are not available for all targets. Dam has a strong propensity to methylate GATC sequences [15, 16]. This is a disadvantage which can cause high background signals. However, methylation marks can persist on the DNA for an extended amount of time and are placed rapidly when Dam is brought into proximity of the DNA. This allows the assay, unlike ChIP, to be utilized in investigations of transient, loose, or indirect interactions [15, 32]. These same properties of the Dam enzyme allow for mapping of three-dimensional chromatin structure [18, 19]. DamID also can be used for global analysis through next-generation sequencing [33].

DamID has some comparative weaknesses, nonetheless. A limitation that must be considered is the potential for the Dam enzyme to interfere with folding of proteins of interest [15]. Positive controls show continued function, but do not guarantee the efficiency of binding or function. As stated above, DamID does have strong background activity requiring effective negative controls and multiple experimental runs. The persistent nature of the methylation means that investigations of interactions under different conditions will continue to show marks from a previously treated state for some time. This can most easily be avoided by extending the time between treatments, allowing DNA replication to drown out the methylation. However, this cannot be applied to all systems such as quiescent and

senescent cells. DamID also does not differentiate between post-transcriptional modifications, where ChIP succeeds [15, 31, 34]. Methylation occurs quickly after Dam is brought into proximity of a GATC sequence but initial expression of the fusion is not immediate upon transduction [15].

My technical triplicates tended to have values close to one another. Outliers were removed by the Dixon Q test with 96% confidence, but often deviated by near inconsequential amounts as low as 0.01 C_t. Biological replicates did not, in general, retain this tight distribution. In particular, any two biological replicates could have vastly different values C_t values. While the magnitude of replicate values differed, signal at positive sites was consistently two-fold higher than the background. Signal at negative control locations or those that did not show evidence of binding remained consistent between replicates. Therefore, it is apparent that DamID can act as a binary test with only a few replicates. Primer sets near a binding site consistently exhibited signals two-fold or greater than the negative controls. However, quantitative binding was difficult to establish. Binding of a single transcription factor at one primer set showed increases ranging from two- to 966-fold over the controls. The large spread in the data led me to use an analysis method more reliant upon a binary threshold (Equation 2).It is possible that the variance between samples originates from the endogenous protein binding in competition with the fusion proteins.

A large portion of the work done here was to establish these important components of the experiment. Nonetheless, once a cell line is established and controls are confirmed, DamID can be run at relatively high throughput. Measuring the binding of a particular Dam fusion to a new target can also be accomplished in a short amount of time.

The negative controls were scattered throughout the genome but provided consistent evidence that DamID was working. Importantly, amplification at the negative control regions was consistently low as assessed by C(n) value (Equations 2, 3) and did not cross the two-fold threshold in most runs. This allowed for accurate identification of binding within unknown sites.

Positive controls were established for use in Dam-CTCF and Dam-TEAD3 assays (Figures 8A, 9A). Strong peaks can be seen in each of the CTCF positive control sites (Figure 8A). CTCF binding is particularly clear in *CDKN2B* where the two primer sets were sufficiently close that their positive binding was additive. These controls worked consistently, showing a two-fold or higher increase over background. Future analyses of CTCF or TEAD3 DNA binding through DamID can use these primer sets as positive controls to compare to other regions. Despite the clear indication of binding through DamID, these positive controls did not exhibit C_t values as consistent as seen in the negative controls. The unfortunate effect of this is that the positive controls cannot be easily used to quantify levels of binding in unknowns or putative regions. The SMAD3 positive control primers were never found to work. Although there is a known SMAD3 interaction site in the *Trangelin* promoter there was no evidence of binding through DamID. Several proposed mechanisms would explain this result. One possibility is that the fusion of Dam to SMAD3 disrupts function of the transcription factor. Alternatively, the Dam-SMAD3 fusion may not be expressed. An immunoblot with an antibody against Dam or SMAD3 could confirm this explanation. Another possibility depends on the activity of SMAD3. Under normal conditions, SMAD3 is localized to both the cytoplasm and nucleus. Signaling initiated by transforming growth factor beta (TGF-β) stimulates nuclear import and the formation of a heteromeric complex with SMAD4. This complex then binds to DNA containing 5′-GTCT-3′ motifs [32]. Therefore, SMAD3 may not have a strong enough interaction with the positive control region in the absence of TGF-β. In the future, DamID could be run after TGF-β treatment to evaluate the validity of this explanation.

The DamID assay, preformed using Dam-CTCF, -TEAD3, and -SMAD3, returned several interesting results. Most notably, it was shown that a putative CTCF binding site in the PRR did indeed display binding. While this does not confirm chromatin looping between the PRR and the *INK/ARF* locus, it does support the possibility. Nonetheless, it demonstrates success of the DamID system in defining novel binding sites. The ECAD4 site also showed strong interaction with TEAD3, suggesting that the downstream putative binding site is utilized (Figure 9B). TEAD3 did show interaction with *CDKN2B,* which would support a mechanism by which it can regulate the *INK4/ARF* locus (Figure 9A). However, binding of TEAD3 was not observed near the transcriptional start sites of $p16^{INK4a}$ or $p14^{ARF}$. Finally, a putative SMAD3 binding site in the PRR did not show evidence of binding. Due to the lack of a good positive control, it is possible my inability to detect SMAD3 binding was an artifact of a non-functional protein. Alternatively, induction of TGF-β signaling may be required before binding is evident. Though these are possibilities, a graduate student in the lab had previously used ChIP-Seq to show that SMAD3 does not bind the PRR (data not shown).

It was difficult to quantitatively compare the interaction of a transcription factor between PCR sites with high confidence due to the large variability between biological replicates. The chosen method depicts relative confidence in interaction and has several useful properties. First, it closely reflects the 1250bp range of DamID because 99% of the peak is contained within that range. Further, it is weighted to have higher confidence in larger data sets with a stronger average amplification (Equation 2). In addition, a smaller standard deviation in the data would return higher confidence (Equation 2). The analysis is also additive such that the use of several proximal primer sets increases confidence that an interaction is real. In the meantime, data from primer sets far away from each are not additive and show no interaction. The additive nature of my analytical design also provides a potential method to narrow the likely window of transcription factor binding. Unfortunately, use of a normal distribution causes the ends of each binding peak to be comparatively small. This takes away from some of the advantages of using an additive analysis. In particular, this method is useful for increasing the confidence in binding sites with many proximal primer sets. On the other hand, primer sets that have a very small overlapping range are not as useful. Alternative analyses could potentially use primer sets with small overlaps to narrow the range of binding for any unknown interaction.

Now that controls are established for CTCF and TEAD3, several new directions are feasible. Firstly, chromatin looping between the *INK/ARF* locus and PRR can be investigated. Transcription activator-like effectors (TALEs) can bind DNA in a sequence specific manner and can be designed for a chosen sequence [35, 36]. TALEs have been effectively fused to several different classes of protein including nucleases and, importantly, epigenetic modifying domains [36]. A TALE-Dam fusion directed specifically to the PRR would help confirm chromatin looping. Only through chromatin looping would the TALE-Dam fusion be brought into proximity of, and methylate the *INK/ARF* locus (Figure 4). To confirm that CTCF is the mediator of this looping, CRISPR/Cas9 endonuclease or another designer nuclease could be used to remove the CTCF site in the PRR. If the *INK/ARF* methylation does not persist, CTCF would be determined as the primary mediator of the event.

The establishment of controls and mapping of steady binding patterns now also allows for investigations of how DNA interactions change under different conditions. In particular, to study OIS, an inducible oncogene (e.g. $NRAS^{Q61R}$) could be inserted into the cell line [37]. By running DamID on both induced and not induced cells, differences in binding could potentially be observed. Alternatively, stable, constitutive expression of an oncogene could be achieved in one cell line which could be compared to a line not expressing the oncogene. Such a study could elucidate the mechanism for

OIS and eventually improve our understanding of how systems to protect against uninhibited cell growth are activated and why they fail.

In sum, I have established DamID controls to profile CTCF and TEAD3 binding to the *INK/ARF* locus. Moving forward, PRR looping with the *INK/ARF* locus could be investigated and the mechanistic details of OIS illuminated.

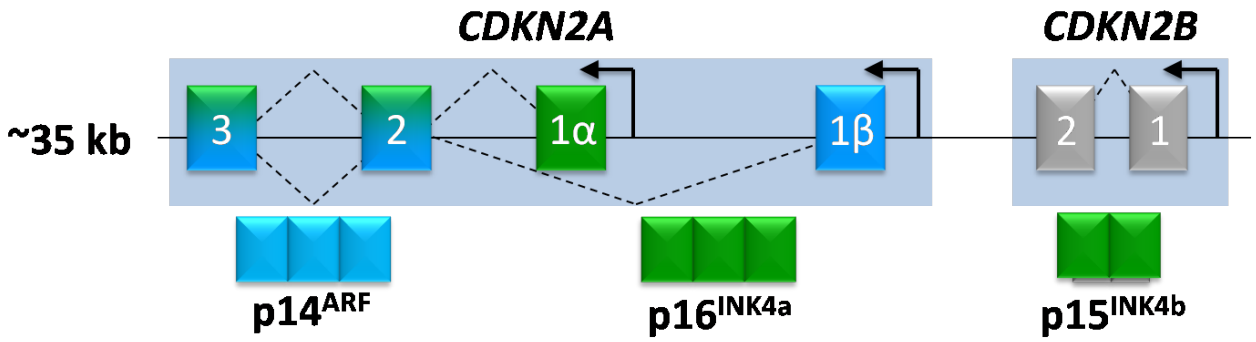


Figure 1: A map of the *INK4/ARF* locus. Located on chromosome 9p21.3, the *INK4/ARF* locus contains encodes two tumor suppressor genes, *CDKN2A* and *CDKN2B*. Exons 2 and 3 of *CDKN2A* are shared by *p14*^{ARF} and *p16*^{INK4a} but each has a different starting exon. *CDKN2B* encodes p15^{INK4b}.

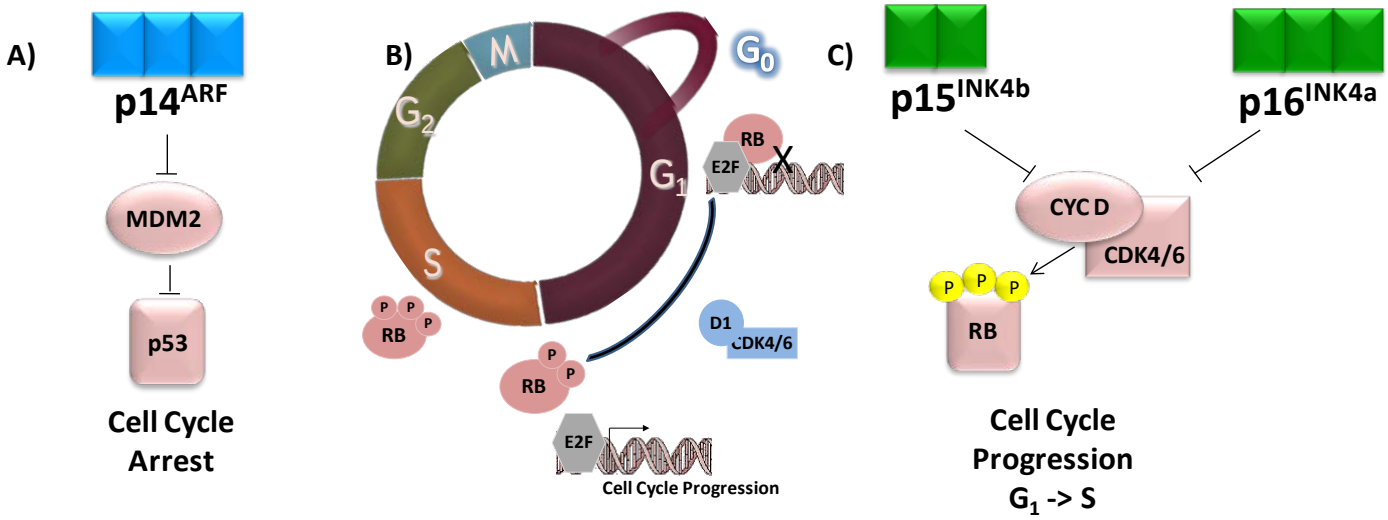


Figure 2: The *INK/ARF* locus controls progression through the cell cycle. A) p14^{ARF} inhibits MDM2 interaction with p53, promoting cell cycle arrest and DNA repair. **B)** Retinoblastoma (Rb) regulates cell cycle progression from G1- to S-phase by binding to, and inhibiting the transcriptional activity of E2F. Phosphorylation of Rb alleviates this interaction, allowing for cell cycle progression. **C)** p15^{INK4b} and p16^{INK4a} inhibit the binding of cyclin D to CDK4/6, thereby preventing Rb phosphorylation and subsequent cell cycle progression.

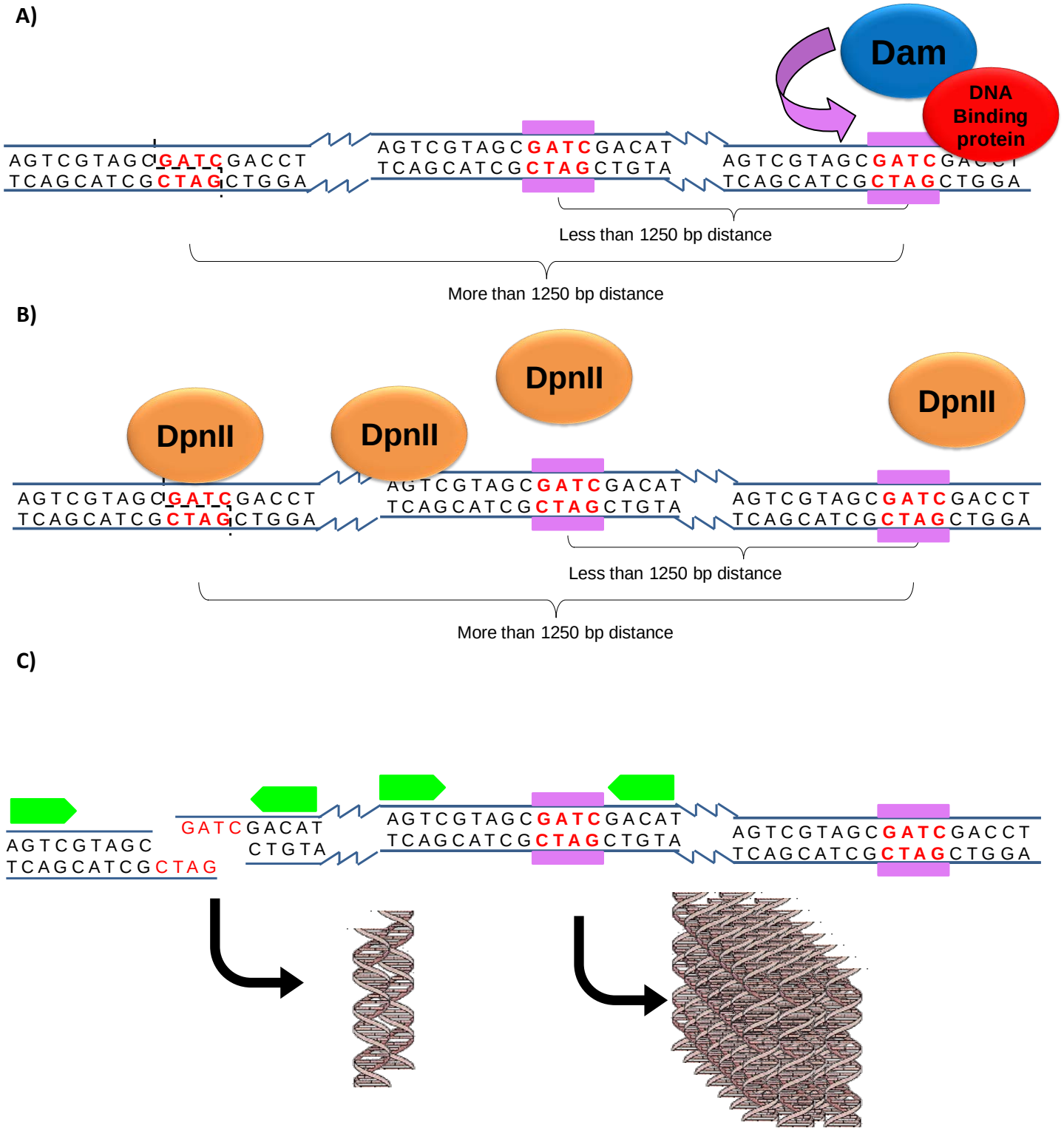


Figure 3: An overview of the DamID assay. Pink bars above the DNA designate a methylated adenine. **A)** The *E.Coli* Dam protein will methylate GATC sequences near the binding sites of the fused DNA-binding factor. **B)** *DpnII* will not cut GATC sequences that are Dam-methylated. **C)** By performing PCR across a potential methylation site (Primers represented by green arrows), one can visualize the loss of digested sequences.

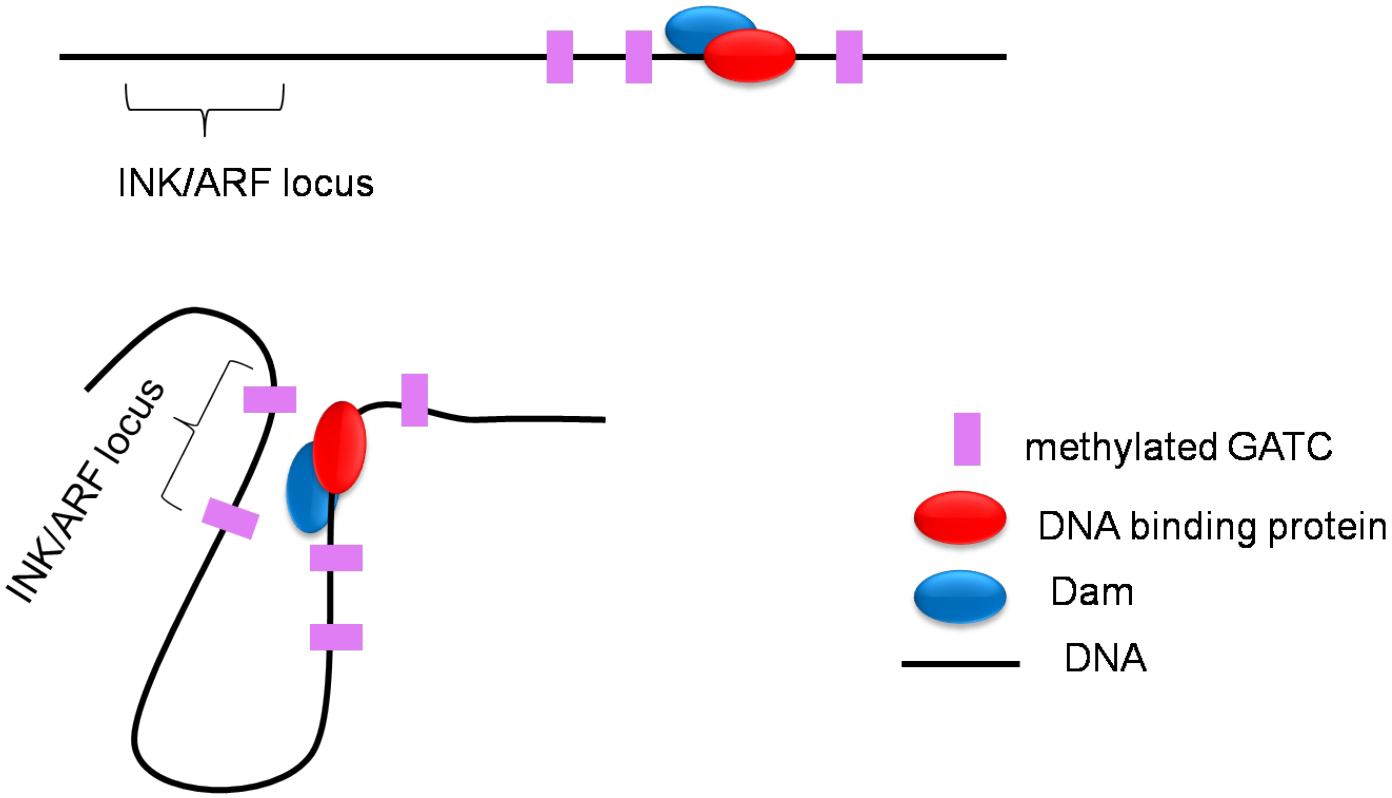


Figure 4: Investigating chromatin looping in using DamID. Linear DNA greater than 1250 bp from the binding site will not be methylated. However, DNA looping can bring additional sites in proximity to the Dam fusion protein.

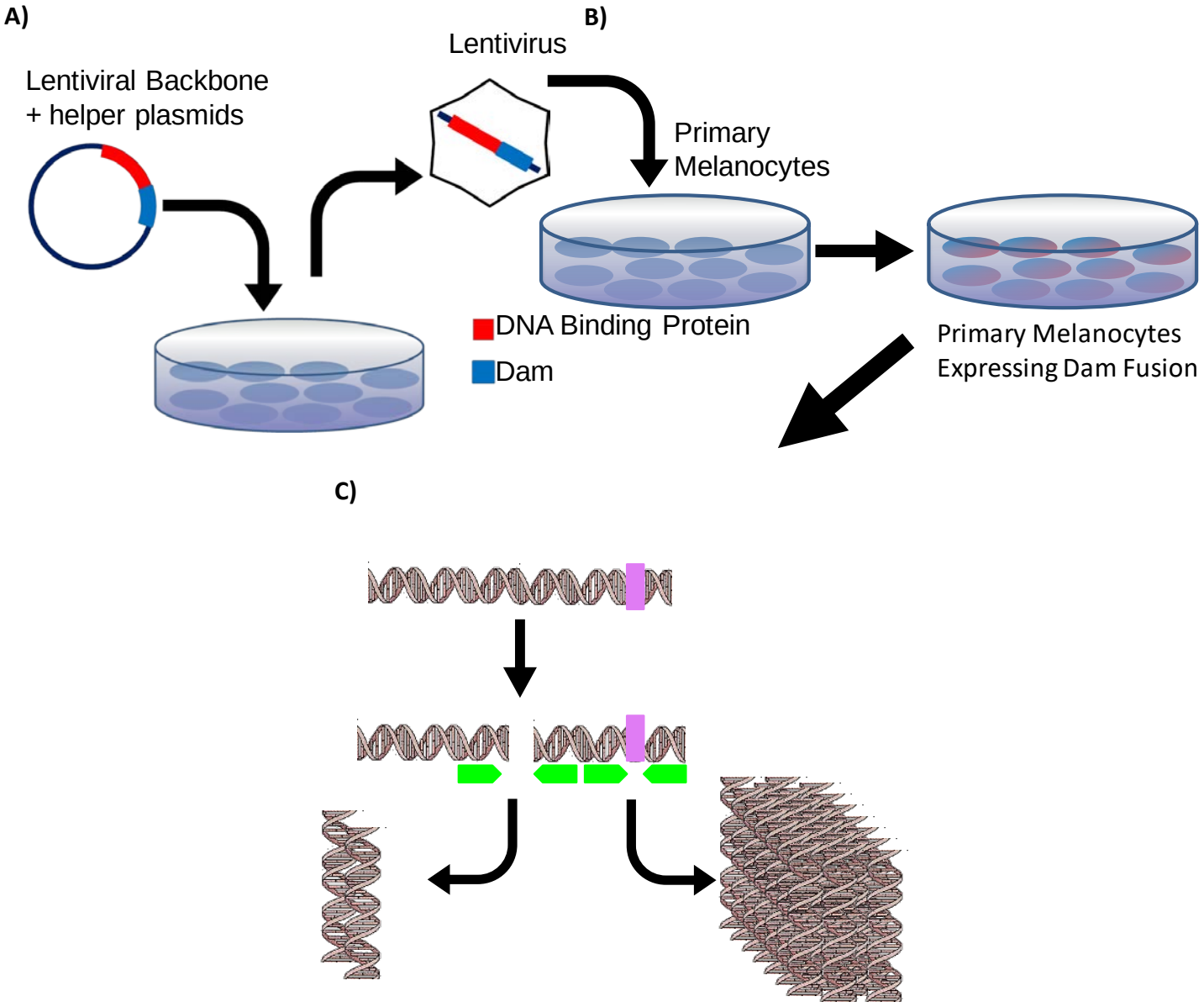


Figure 5: An overview of the DamID procedure. Pink bars above the DNA designate a methylated adenine. **A)** Lentiviral backbones are created that encode Dam-transcription factor fusions. Each backbone is transfected into 293t packaging cells with helper plasmids encoding viral structual proteins. Mature lentivirus is harvested from the media. **B)** Lentiviruses are placed on human primary melanocytes for infection. Hygromycin selection is performed to kill off any uninfected cells. **C)** DNA is isolated from melanocytes stably expressing the lentiviral constructs and digested with *DpnII*. SYBR-based qPCR is performed around GATC sequences of interest to compare methylation status near potential binding sites of interest.

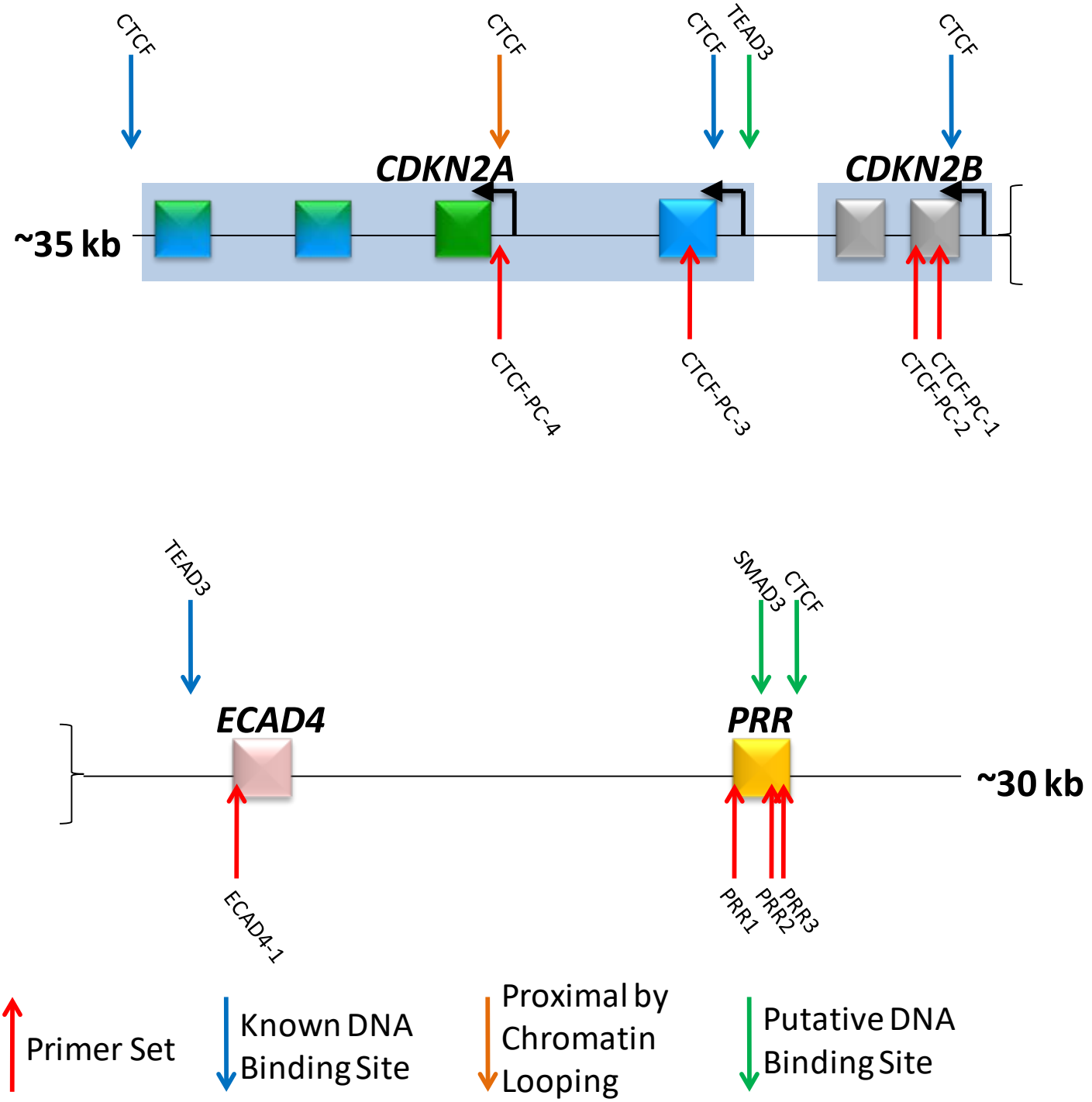
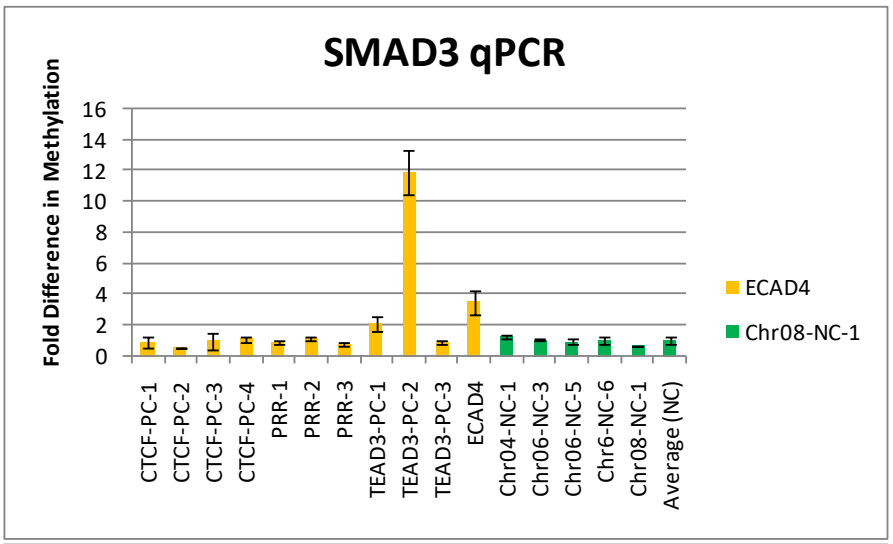
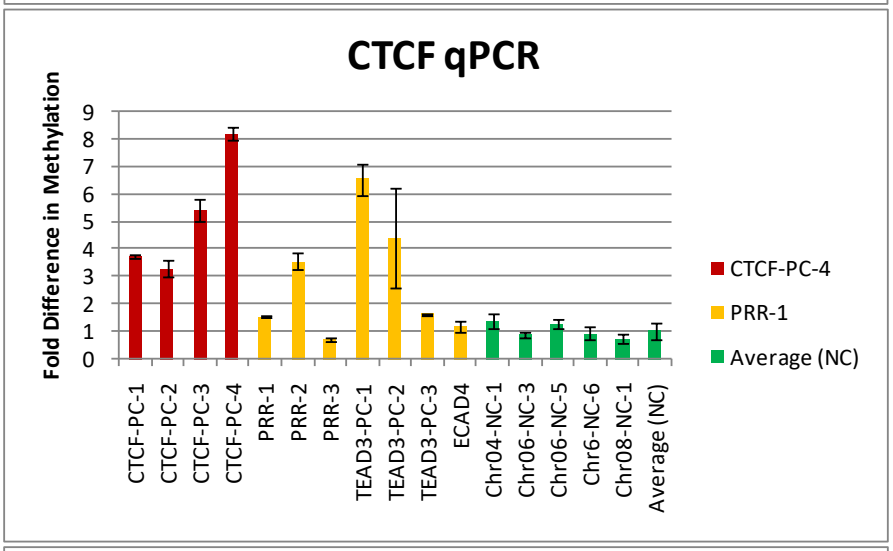


Figure 6: The relative location of known and putative DNA-binding sites as well as primer sets in the *INK/ARF* locus and regulatory region. Distances are not to scale. There is a distance of about 60kb between A and B. **A)** Primer sets in the *INK/ARF* locus with key features. **B)** Primer sets in the *INK/ARF* regulatory region.

A)



B)



C)

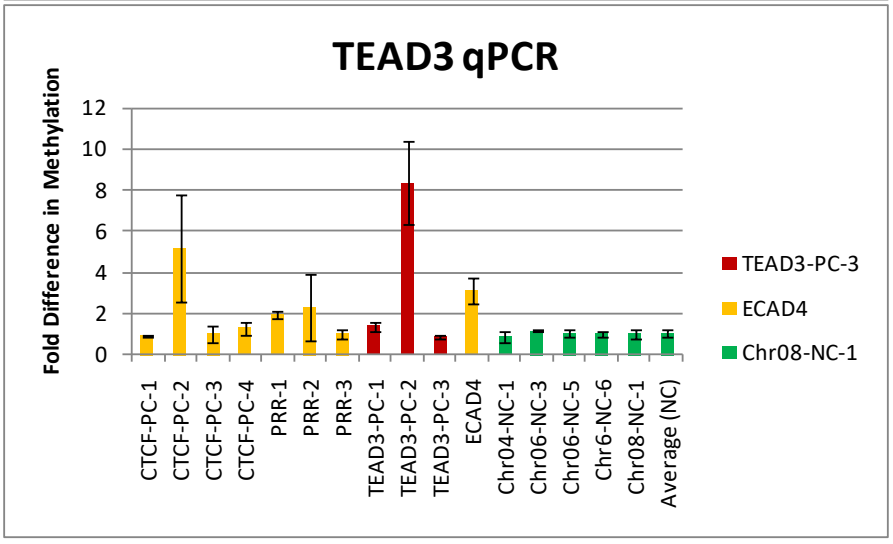


Figure 7: Representative DamID results. Shown is agraph of therelative DNA abundance at each site as detected by qRT-PCR and normalized to the average value of each respective negative control. Representative results are shown for Dam-SMAD (**A**),Dam-CTCF (**B**), and Dam-TEAD3 (**C**).

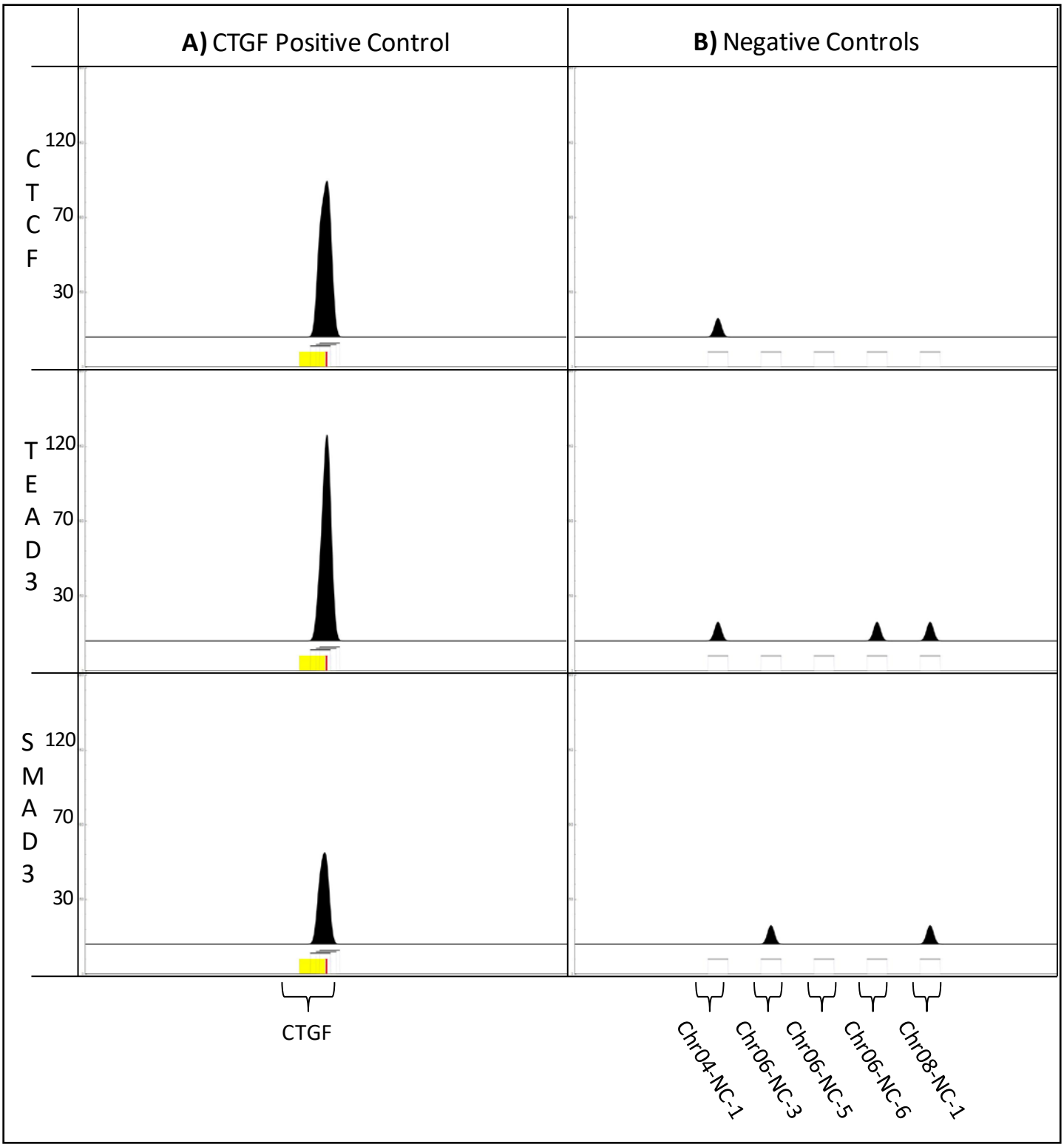


Figure 8: Relative binding of CTCF, SMAD, and TEAD3 around *CTGF* (positive control) and negative controls. Peaks, centered at the assayed GATC sequence, are graphed as a normal distribution with standard deveations equal to 1/3 the range of Dam methylation. The peak height represents relative confidence in binding as described in the methods. All images are similalry scaled. **A)** Binding in the proximity of the *CTGF* positive control. **B)** Binding detected within negative control primer sets. Although they are represented together here, the negative controls are not in proximity to each other in the genome.

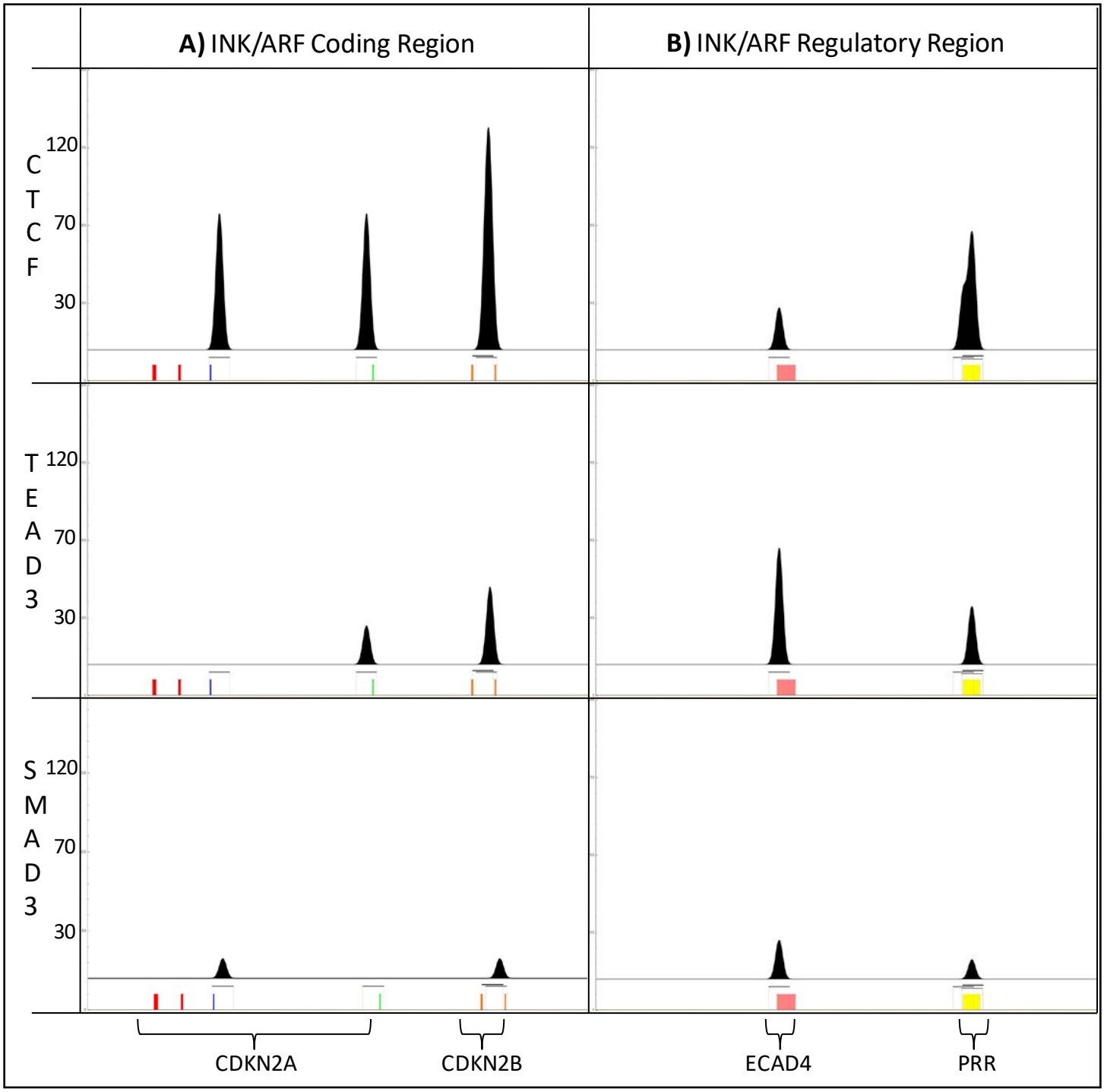


Figure 9: Relative binding of CTCF, SMAD, and TEAD3 within the *INK/ARF* locusand regulatory region. Peaks, centered at the assayed GATC sequence, are graphed as a normal distribution with standard deveations equal to 1/3 the range of Dam methylation. The peak height represents relative confidence in binding as described in the methods. All images are scaled in the same manner. **A)** Detected binding in the proximity of *INK/ARF* locus. **B)** Detected binding at the regulatory region approximately 90kb upstream of the *INK/ARF* locus.

Primer	Sequence (5'->3')	Location in hg38
Chr04-NC-1 F	TCACTTTCAGACACTGATGTTG	Chr4:93587020-93587041
Chr04-NC-1 R	GAACAGAGGGAATAAAGAGAAATGATAG	Chr4:93587128-93587101
Chr06-NC-3 F	GTCATTCCATGGGTACTAACCT	Chr6:12984638-12984659
Chr06-NC-3 R	TATGAAACAAACTGCTCGGTTG	Chr6:12984782-12984761
Chr06-NC-5 F	CTTGTGAGGACGGGTGTATG	Chr6:12986740-12986759
Chr06-NC-5 R	ACTTCTCTAAACCTTTGCTTACCT	Chr6:12986880-12986857
Chr06-NC-6 F	AAGGCAGCTTCTAGTTCCAAA	Chr6:12987543-12987563
Chr06-NC-6 R	CTGAATGAGCTTTATTTGCTCAAGA	Chr6:12987657-12987633
Chr08-NC-1 F	TGTCTCTCATTGTAAATTGTTGCAC	Chr8:64039557-64039581
Chr08-NC-1 R	TAGAAAGTCAGATGCCATTACAGAA	Chr8:64039751-64039727
CTCF-PC-1 F	GCAGGATTCGTACTTAAACATTGA	Chr9:22008067-22008090
CTCF-PC-1 R	CATACTCAGTGCCCAGATTACA	Chr9:22008206-22008185
CTCF-PC-2 F	CAATTCAGTCTATTCCTTGCATCTC	Chr9:22008534-22008558
CTCF-PC-2 R	CGGAGGTGTGCATTCCAC	Chr9:22008679-22008662
CTCF-PC-3 F	GACCTCTACCTCTAACTCACAAAG	Chr9:21993503-21993526
CTCF-PC-3 R	CGACTTCCTGAAATGCTAACAAG	Chr9:21993630-21993608
CTCF-PC-4 F	CGGACTCCATTCTCAAAGTCATA	Chr9:21975740-21975762
CTCF-PC-4 R	GTGAAGGAGACAGGACAGTATTT	Chr9:21975854-21975832
TEAD3-PC-1 F	GTCCTACACAAACAGGGACAT	Chr6:13195144-131951464
TEAD3-PC-1 R	GGAGGAATGCTGAGTGTCAA	Chr6:131951544-131951525
TEAD3-PC-2 F	CAGTCCGAGCGGTTTCTTT	Chr6:131950682-131950700
TEAD3-PC-2 R	CAAGGGCCTATTCTGTCACTTC	Chr6:131950828-131950807
TEAD3-PC-3 F	CCTCAAGATGCCTACCTGTAAA	Chr6:131951917-131951938
TEAD3-PC-3 R	CGCGTCTTTGTTCTCTTTCTTG	Chr6:131952018-131951997
ECAD4-1 F	TCTGCTTGCCTTCGTAACC	Chr9:22096862-22096880
ECAD4-1 R	GCAAATCCAGCAGGCAAAG	Chr9:22096973-22096955
PRR-1 F	CTGCCTATCTGACCATTGTACTT	Chr9:22118951-22118973
PRR-1 R	TGATAGCATAGTGATTCATACTCCAG	Chr9:22119049-22119024
PRR-2 F	AACTTGAGCTTGGGTTCAG	Chr9:22119947-22119965
PRR-2 R	CCCATATTTCAGACATAACTTTCTC	Chr9:22120062-22120038
PRR-3 F	TGTGAAGATTCAATGAGTTGTAACG	Chr9:22120087-22120111
PRR-3 R	GCTGTCTGTAAGATACTGGGAAG	Chr9:22120214-22120192

Table 1: A list of primers used for DamID qPCR.

References

1. Lapak, K.M. and C.E. Burd, *The Molecular Balancing Act of p16INK4a in Cancer and Aging.* Mol Cancer Res, 2013.
2. Khan, S.A., et al., *p53 Mutations in human cholangiocarcinoma: a review.* Liver Int, 2005. **25**(4): p. 704-16.
3. Kusy, S., C.J. Larsen, and J. Roche, *p14ARF, p15INK4b and p16INK4a methylation status in chronic myelogenous leukemia.* Leuk Lymphoma, 2004. **45**(10): p. 1989-94.
4. Thwaites, M.J., Checchini M.J., Dick, F.A., *Analyzing RB and E2F During the G1–S Transition*, in *Cell Cycle Control. Methods in Molecular Biology (Methods and Protocols)*. 2014, Humana Press: New York, NY.
5. Peeper, D.S., *Oncogene-induced senescence and melanoma: where do we stand?* Pigment Cell Melanoma Res, 2011. **24**(6): p. 1107-11.
6. Campisi, J., *Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors.* Cell, 2005. **120**(4): p. 513-522.
7. Harismendy, O., et al., *9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response.* Nature, 2011. **470**(7333): p. 264-8.
8. Kumar, J., et al., *Association of polymorphisms in 9p21 region with CAD in North Indian population: replication of SNPs identified through GWAS.* Clin Genet, 2011. **79**(6): p. 588-93.
9. Maccioni, L., et al., *Variants at the 9p21 locus and melanoma risk.* BMC Cancer, 2013. **13**: p. 325.
10. Jarinova, O., et al., *Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus.* Arterioscler Thromb Vasc Biol, 2009. **29**(10): p. 1671-7.
11. Hirosue, A., et al., *Quantitative assessment of higher-order chromatin structure of the INK4/ARF locus in human senescent cells.* Aging Cell, 2012. **11**(3): p. 553-6.
12. Witcher, M. and B.M. Emerson, *Epigenetic silencing of the p16(INK4a) tumor suppressor is associated with loss of CTCF binding and a chromatin boundary.* Mol Cell, 2009. **34**(3): p. 271-84.
13. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-45.
14. de Wit, E., et al., *CTCF Binding Polarity Determines Chromatin Looping.* Mol Cell, 2015. **60**(4): p. 676-84.
15. Greil, F., C. Moorman, and B. van Steensel, *DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase.* Methods Enzymol, 2006. **410**: p. 342-59.
16. van Steensel, B. and S. Henikoff, *Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase.* Nat Biotechnol, 2000. **18**(4): p. 424-8.
17. Barras, F. and M.G. Marinus, *The great GATC: DNA methylation in E. coli.* Trends Genet, 1989. **5**(5): p. 139-43.
18. Lebrun, E., et al., *A methyltransferase targeting assay reveals silencer-telomere interactions in budding yeast.* Mol Cell Biol, 2003. **23**(5): p. 1498-508.
19. Cleard, F., et al., *Probing long-distance regulatory interactions in the Drosophila melanogaster bithorax complex using Dam identification.* Nat Genet, 2006. **38**(8): p. 931-5.
20. Benskey, M., Manfredsson, F.P., *Lentivirus Production and Purification*, in *Gene Therapy for Neurological Disorders. Methods in Molecular Biology*. 2016, Humana Press: New York, NY.
21. Garraway, L.A., et al., *Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma.* Nature, 2005. **436**(7047): p. 117-22.
22. Sambrook, J. and D.W. Russell, *Purification of nucleic acids by extraction with phenol:chloroform.* CSH Protoc, 2006. **2006**(1).
23. Salkind, N.J. and K. Rasmussen, *Encyclopedia of measurement and statistics*. 2007, Thousand Oaks, Calif.: SAGE Publications.
24. Efstathiou, C.E., *Stochastic Calculation of Critical Q-Test Values for the Detection of Outliers in Measurements.* Journal of Chemical Education, 1992. **69**(9): p. 733-736.
25. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2016. **44**(D1): p. D7-19.
26. Almontashiri, N.A., et al., *9p21.3 Coronary Artery Disease Risk Variants Disrupt TEAD Transcription Factor-Dependent Transforming Growth Factor beta Regulation of p16 Expression in Human Aortic Smooth Muscle Cells.* Circulation, 2015. **132**(21): p. 1969-78.

27. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
28. Aldeiri, B., et al., *Abrogation of TGF-beta signalling in TAGLN expressing cells recapitulates Pentalogy of Cantrell in the mouse.* Sci Rep, 2018. **8**(1): p. 3658.
29. Zhao, B., et al., *TEAD mediates YAP-dependent gene induction and growth control.* Genes Dev, 2008. **22**(14): p. 1962-71.
30. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
31. Zecchini, V. and I.G. Mills, *Putting chromatin immunoprecipitation into context.* J Cell Biochem, 2009. **107**(1): p. 19-29.
32. Bianchi-Frias, D., et al., *Hairy transcriptional repression targets and cofactor recruitment in Drosophila.* PLoS Biol, 2004. **2**(7): p. E178.
33. Vogel, M.J., D. Peric-Hupkes, and B. van Steensel, *Detection of in vivo protein-DNA interactions using DamID in mammalian cells.* Nat Protoc, 2007. **2**(6): p. 1467-78.
34. Collas, P., *The current state of chromatin immunoprecipitation.* Mol Biotechnol, 2010. **45**(1): p. 87-100.
35. Bochtler, M., *Structural basis of the TAL effector-DNA interaction.* Biol Chem, 2012. **393**(10): p. 1055-66.
36. Kühn, R., W. Wurst, and B. Wefers, *TALENs: Methods and protocols*. 2015, New York: Humana Press.
37. Gos, A., et al., *Molecular characterization and patient outcome of melanoma nodal metastases and an unknown primary site.* Ann Surg Oncol, 2014. **21**(13): p. 4317-23.